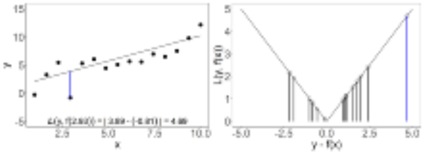
LOSS

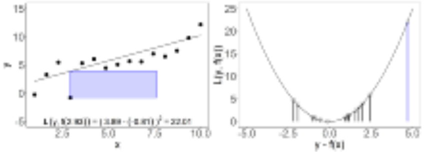
The **loss function** $L(y, f(\mathbf{x}))$ quantifies the "quality" of the prediction $f(\mathbf{x})$ of a single observation $\mathbf{x}$:

$$L : \mathcal{Y} \times \mathbb{R}^g \to \mathbb{R}.$$

In regression, we could use the absolute loss $L(y, f(\mathbf{x})) = |f(\mathbf{x}) - y|$;



or the L2-loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$:



RISK OF A MODEL / 2

Problem: Minimizing $\mathcal{R}(f)$ over f is not feasible:

- $\mathbb{P}_{xy}$ is unknown (otherwise we could use it to construct optimal predictions).
- We could estimate $\mathbb{P}_{xy}$ in non-parametric fashion from the data $\mathcal{D}$, e.g., by kernel density estimation, but this really does not scale to higher dimensions (see "curse of dimensionality").
- We can efficiently estimate $\mathbb{P}_{xy}$, if we place rigorous assumptions on its distributional form, and methods like discriminant analysis work exactly this way.

But as we have n i.i.d. data points from $\mathbb{P}_{xy}$ available we can simply approximate the expected risk by computing it on $\mathcal{D}$.

EMPIRICAL RISK / 2

- The risk can also be defined as an average loss

$$\bar{\mathcal{R}}_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right).$$

The factor $\frac{1}{n}$ does not make a difference in optimization, so we will consider $\mathcal{R}_{\text{emp}}(f)$ most of the time.

- Since f is usually defined by **parameters** θ, this becomes:

$$\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$$

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

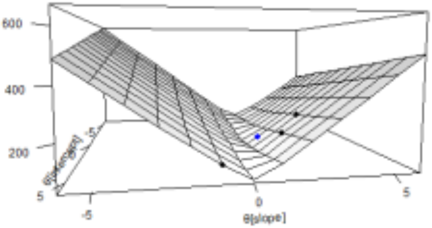
EMPIRICAL RISK MINIMIZATION

But usually $\mathcal{H}$ is infinitely large.

Instead we can consider the risk surface w.r.t. the parameters $\boldsymbol{\theta}$.
(By this I simply mean the visualization of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$)

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}.$$

Model	$\theta_{intercept}$	θ_{slope}	$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
f_1	2	3	194.62
f_2	3	2	127.12
f_3	6	-1	95.81
f_4	1	1.5	57.96



EMPIRICAL RISK MINIMIZATION / 2

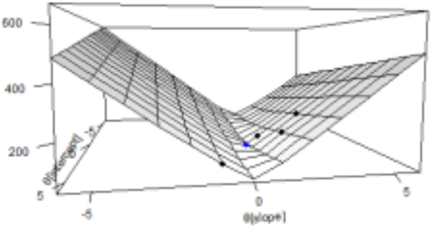
Minimizing this surface is called **empirical risk minimization** (ERM).

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}).$$

Usually we do this by numerical optimization.

$$\mathcal{R} : \mathbb{R}^d \to \mathbb{R}.$$

Model	$\theta_{intercept}$	θ_{slope}	$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
f_1	2	3	194.62
f_2	3	2	127.12
f_3	6	-1	95.81
f_4	1	1.5	57.96
f_5	1.25	0.90	23.40



In a certain sense, we have now reduced the problem of learning to **numerical parameter optimization**.